EXHIBIT B

Voice Recognition in Cellular Mobile Telephones

A speaker-independent system has been developed which allows "hands-free" telephone use.

Thomas B. Schalk
Director of Technology Development
Voice Control Systems
Dallas, TX

The advent of cellular telephones marked the beginning of high quality mobile telephone service. The first mobile cellular service was offered in Chicago during 1983 and now over 100 major cities offer the service. Accompanying this, has been the proliferation of 500,000 mobile telephone users, all of whom face the challenge of manipulating their phones in a mobile environment.

A typical cellular telephone has approximately 20 keys that correspond to the 10 digits and various control functions. Dialing telephone numbers while driving can be dangerous because the user will typically take his or her eyes off the road to manipulate the telephone keypad. Even after the phone number has been dialed, the user must hold the handset which makes shifting gears and using turn signals, among other things, difficult to do. Hence, many cellular phones have an optional remote microphone that is mounted near the visor, and a speaker located somewhere near the driver. After a call has been placed, the remote microphone and speaker are used in a "hands-free" fashion.

The voice-dialing mobile cellular telephone is one of the most exciting and promising applications of speech recognition in telephony. The use of voice input for dialing can alleviate many of the safety problems associated with cellular telephone systems. A speaker-independent voice recognition system for cellular phones has been developed. This voice control unit is designed to operate optimally in driving vehicles. To develop this system, an extensive voice data collection took place.

In this article the performance requirements for this application will be considered first, then data collection procedures and the functional capabilities of the recognizer will be discussed.

Recognition Performance Requirements

The recognition technology used in the voice control unit for the cellular telephone is speaker-independent and operates on isolated speech. For operating the phone by voice, in place of the key pad, only a small fixed vocabulary is needed. The primary advantage of using speaker-independent technology for the cellular application is that the recognizer need not be trained by each user. Since the noise encountered during driving conditions varies tremendously, it is not practical to expect users to train recognizers properly for the mobile environment. For example, should the user train the system with the engine on or off? Should the blower be set at low, medium, or high? On what road surfaces should the user drive during training, and at what speeds? The speaker-independent reference data used for development of the voice control unit were collected under a wide variety of conditions, which accounts for the robustness of the resultant recognizer in various mobile environmental noises.

It should be noted, though, that a speaker-dependent capability would be useful because it offers vocabulary flexibility. This would allow the user to create custom vocabularies that include people's names to facilitate speed dialing. For example, a person could simply pick up the handset and say "speed-dial Bob Smith" and the preprogrammed phone number for Bob Smith would be dialed automatically.

Commercially available speech recognizers exhibit a wide range of recognition performance. To assess performance, recognition accuracy of the system must be measured. In a crude sense, recognition accuracy refers to the percentage of the time that the recognizer correctly classifies an input utterance. It depends on a number of factors, such as whether the system is speaker-dependent or speaker-independent, whether the system is a discrete or connected word recognizer, the difficulty of the vocabulary, the cost of the system, and the environment in which the system is used. The most stringent performance requirement for the cellular application is reliable digit recognition. Since typical phone numbers are 7 digits long, individual digit accuracy must be very high to make dialing phone numbers by voice practical. Therefore, the ability to detect recognition errors and correct them is critical.

There are three types of errors that a recognizer can make. One, the most obnoxious, is called a substitution error. A substitution occurs when an incorrect word is hypothesized for a valid input utterance. For example, if the active vocabulary for a recognizer includes digits and a "two" is hypothesized when a "nine" was actually spoken, then the recognizer is said to have substituted a two for a nine. In general, substitution error rates must be less than 2 percent for user acceptance. The speaker-independent technology developed for the cellular telephone application achieves this performance goal based on measurements from a large data base of speakers who were "naive" and "inexperienced" recognizer users. The "experi-

Voice Dialing can alleviate many of the safety problems of cellular telephone use.

enced" user tends to incur much lower error rates, making the successful recognition of seven consecutive digits quite likely.

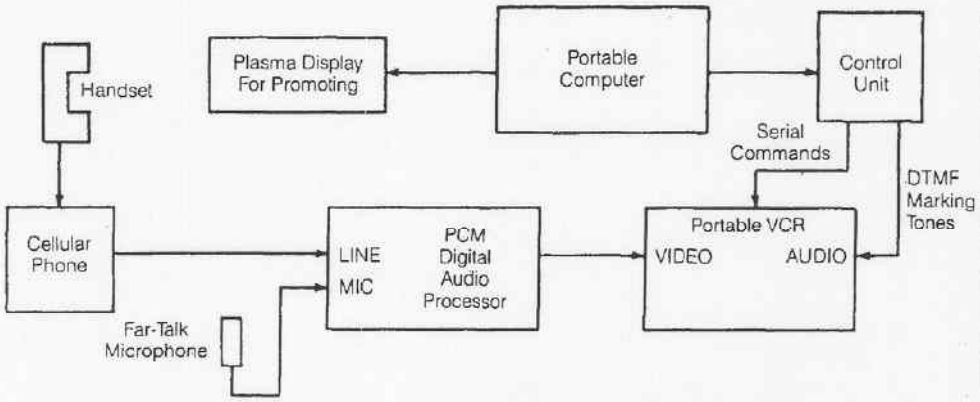
The second type of recognition error is a rejection error. This occurs when a valid input utterance is not classified by the recognizer. When rejections occur, the user simply repeats the utterance—ideally one time—until it is recognized. Rejection errors are not as obnoxious as substitutions, but should not occur more than 5 percent the time for user acceptance.

The third type of recognition error is a spurious response error. This occurs when an invalid input "sound" (such as a horn honking or uttering a word not found in the vocabulary) is classified as a vocabulary word. Ideally, a recognizer should reject all spurious input. Unfortunately, none of today's recognizers are immune to it. Spurious responses can be minimized by either using push-to-talk microphone arrangements or close-talk microphones. The cellular phone recognition system described here was designed to work under close-talk conditions with a cellular handset. An experimental system using a "far-talk" microphone has also been developed and it uses a push-to-talk microphone activation system to minimize spurious errors.

It is not feasible to quantitatively measure the spurious response error rate for a given recognizer. To do so would involve collecting a data base of all possible sounds that can occur. Nevertheless, if such a data base did exist, spurious response error rates of 50 percent would not be surprising for a typical recognizer. This means that about half the time a spurious sound occurs that is loud enough, the recognizer will attempt to classify it as a word in the vocabulary.

Fig. 1

DATA COLLECTION SYSTEM FOR MOBILE ENVIRONMENT



This data collection process took about 10 minutes per individual.

Data Collection Procedure

Speech recognizers generally exhibit sensitivity to changes in the environments in which they are used. The noise characteristics of moving vehicles vary dramatically depending on the driving conditions and the characteristics of the vehicle itself—say a luxury compared to an inexpensive small car. To develop the speaker-independent reference data for the voice control unit, a large data-base collection was conducted in moving vehicles. Data samples were collected from over 500 people—approximately 100 speakers each in 5 different automobiles selected to span a wide range of driving noise. For safety reasons, the voice donor sat on the passenger side during data collection.

There were two phases to the data collection. The first phase involved the collection of data from a remote ("far-talk") microphone mounted near the visor area. During the second phase, samples were collected through a cellular telephone handset. For each phase of the collection, the voice donor was instructed to say words as they appeared on a custom built prompting display located on the passenger's side of the dashboard. The voice donors were instructed to say the words quickly, in an authoritative manner.

Immediately after the far-talk portion of the collection was completed, the volunteer was again instructed to speak the words as they appeared on the monitor, but this time they were to be said into the cellular handset as though they were "conversing" with it. The entire data collection process took approximately 10 minutes per individual.

The speech samples were recorded using portable digital audio processing recording equipment. This equipment consisted of a SONY Portable VCR SI-2000 and a SONY PCM F-1 Digital Audio Processor (Fig. 1). A small portable computer was used to control the recording equipment and was programmed to feed the speech vocabulary prompts to the prompting display located on the passenger side of the dashboard. In addition, the tapes were automatically marked to indicate speaker number, speech track boundaries, and prompting information. Since only the video portion of the tape was used for PCM speech recording, the normal audio track was available for writing ASCII coded DTMF tones to code this information. This allowed for unsupervised digitization of the speech data onto the VAX computer for speaker-independent vocabulary development. The vocabulary words collected through the handsets included the following list of words:

1. one
2. two
3. three
4. four
5. five
6. six
7. seven
8. eight
9. nine
10. zero
11. oh
12. dial
13. recall
14. speed-dial
15. send
16. cancel
17. clear
18. verify
19. spouse
20. home
21. friend
22. work
23. office
24. school
25. service
26. information
27. airline
28. emergency

During the data-base collection, an emphasis was placed on obtaining a reasonable distribution of different dialects, and on collecting an equal number of male and female voices. As stated earlier, five different cars were used in the collection, as well as a variety of handsets. Most of the speech was collected in a moving vehicle under many different environmental conditions. Some of these conditions were rain with windshield wipers on, defroster on, air conditioner set at various levels, heater set at various levels, windows open, and a ra-

dio playing in the background. Furthermore, the data samples were collected in many different locations throughout the Dallas/Ft. Worth Metroplex yielding driving conditions that differed greatly from one location to the next. Prior to the development of the speaker-independent reference data, all of the collected speech was audited and documented. Special attention was given to dialect variations, age of speaker, and any extraordinary characteristics associated with either the speaker or the driving conditions at the time of the collection.

Cellular Telephone Voice Control Unit

The cellular telephone voice control unit discussed here has some noteworthy characteristics. The voice recognition technology is speaker-independent, thus there is no user training required. This specialized recognition system was designed specifically for the noise characteristics of the mobile communications environment. It is a tware-based recognizer that requires a single general purpose microprocessor (Intel 80186) for implementation. The recognition circuit interfaces to the mobile telephone through the bus that connects the phone control unit (located near the vehicle driver) to the transceiver unit (mounted in the vehicle trunk) as indicated in Fig. 2.

The functional operation of the voice unit centers around syntactically structured voice commands from the user, and voice responses from the voice control unit. The command syntax structure is illustrated in Fig. 3. This simple scheme for voice dialing involves a recognition vocabulary of only 28 words (listed earlier). The output channel of the CODEC in the voice recognizer front-end is used for voice responses to guide the user and provide aural feedback for validation of input. As illustrated in Fig. 2, the voice control unit taps into the voice channel-and-control interface on the control bus of the cellular telephone system. The voice control unit recognizes voice commands given to the phone and then issues appropriate commands to operate the telephone.

Each voice command to the phone is acknowledged by the voice control unit through an aural response. If the command is recognized, then a short beep-tone signifies to the user that the voice control unit has recognized the command and is ready for the next one. If the measured signal-to-noise ratio of a detected utterance is below 20 dB, the utterance is rejected and the voice control unit asks the user to "speak louder." If an utterance has an acceptable signal-to-noise ratio, but is not identified with sufficient confidence, it is rejected, and the response "repeat" is issued to the user, indicating he or she should re-enter the command.

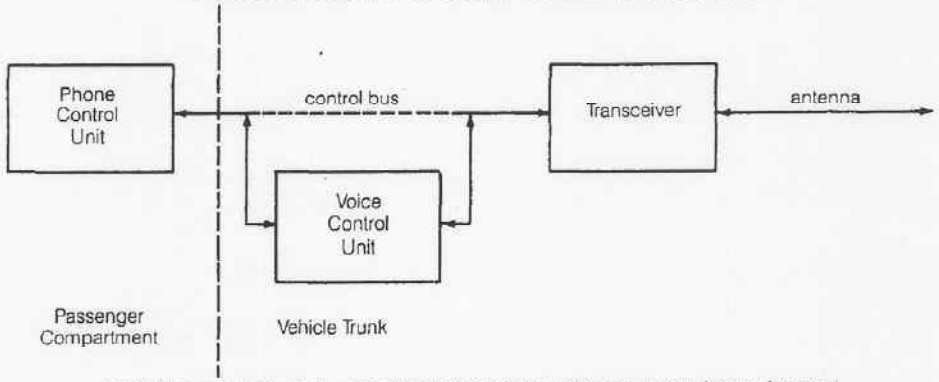
To dial phone numbers, the user simply says "dial" followed by a string of digits. After speaking the last digit of the telephone number, the user says "verify" and a voice response system is activated which repeats the recognized digit sequence through the earpiece of the handset. The user then says "send" in order for that number to be dialed, or "clear" if an improper digit sequence occurred, which activates the top-node vocabulary (dial, recall, speed-dial) and "ready" is synthesized.

For memory dialing (preprogrammed phone numbers), the user says "recall" and then utters a one or two digit sequence, depending on how many preprogrammed numbers the cellular telephone can store. The recognized sequence is repeated and the corresponding phone number in memory is dialed if the word "send" is spoken and recognized. Speed-dialing is acheived by simply picking up the handset and saying "speed dial" followed by one of the ten destination descriptors such as "home," "office," "friend," etc. The recognized words are then repeated and the user can activate the call by saying "send."

Enhanced Cellular Telephone Use

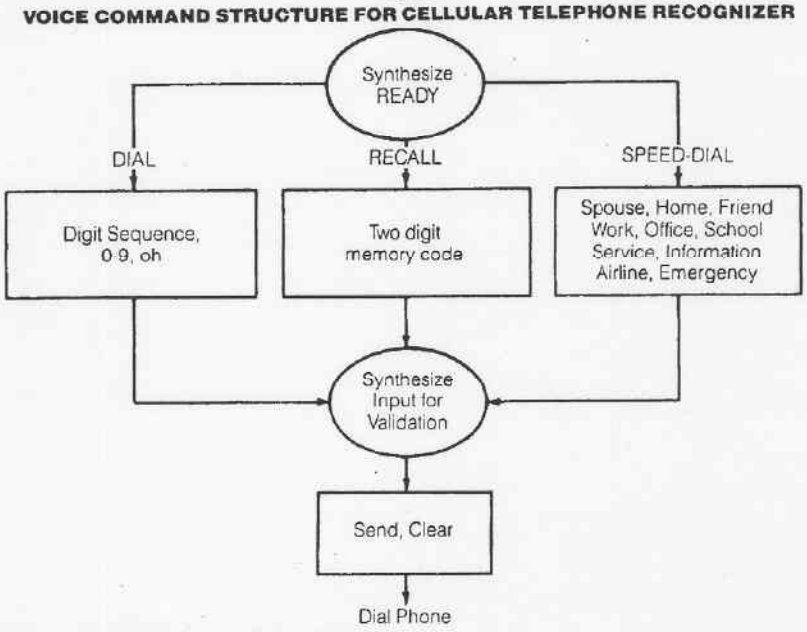
The cellular telephone voice control unit

Fig. 2
VOICE-CONTROLLED MOBILE CELLULAR TELEPHONE



It was designed specifically for noise characteristics of the mobile communications environment.

Fig. 3
VOICE COMMAND STRUCTURE FOR CELLULAR TELEPHONE RECOGNIZER



A 28-word vocabulary is used for this voice-dialing system.

achieves robust speaker-independent speech recognition in a highly variable high-noise environment. Performance measurements using the data base collected in the mobile automotive environment yielded substitution error below 2 percent and rejection rates below 3 percent. The system tested has been implemented in relatively simple hardware which is a fraction of the cost of the cellular telephone unit it enhances. The cellular telephone voice control unit stands as one of the best examples of a fruitful application of speech recognition technology. It achieves its high practical value by simplifying a critical interface and thus significantly enhancing the safety of cellular telephone use.

FOR MORE INFORMATION

Contact Thomas B. Schalk, Voice Control Systems, 14140 Midway Road, Suite 100, Dallas, TX 75244. (214) 386-0300.

THOMAS B. SCHALK *is Director of Technology Development at Voice Control Systems. Currently, he directs the VCS research staff, which is working towards advancing speaker-independent speech recognition technology. This effort has led to new technologies that include a phonetic approach for speaker-independent recognition of isolated speech. Prior to joining VCS, he was employed at Texas Instruments, where he conducted speech research and managed a large government contract to develop speech technology.*

Dr. Schalk is experienced in speech data base design and collection precedures and has patents pending for developing speaker-dependent and speaker-independent recognition technology. Dr. Schalk earned a B.S. in Electrical Engineering from George Washington University and received a Ph.D. in Auditory Physiology from the Johns Hopkins School of Medicine.